Fuzzy c-means based coincidental link filtering in support of inferring social networks from spatiotemporal data streams

Zhang, Pu; Shen, Qiang

Published in:
Soft Computing

DOI:
10.1007/s00500-018-3363-y

Publication date:
2018

Citation for published version (APA):
Zhang, P., & Shen, Q. (2018). Fuzzy c-means based coincidental link filtering in support of inferring social networks from spatiotemporal data streams. *Soft Computing*, *22*(21), 7015-7025. https://doi.org/10.1007/s00500-018-3363-y

Soft Computing
https://doi.org/10.1007/s00500-018-3363-y

FOCUS

Fuzzy c-means based coincidental link filtering in support of inferring social networks from spatiotemporal data streams

Pu Zhang[1] · **Qiang Shen**[1]

Abstract
Social network modelling offers an important computational mechanism for analysis of complex system behaviour. Social networks can be established by linking individuals based on observations of certain activities like physical proximity, by exploiting spatiotemporal data streams that have been acquired. A potentially powerful approach to building such networks is to computationally imitate the real-world foraging process of great tits, where the data streams required consist of the times and locations each tit of a certain population has appeared at. The method works by clustering individuals within the population into groups representing different gathering events with respect to the time and location. It links up the individuals appearing in the same events and subsequently, filters out those links that are set up coincidentally. However, the original filtering technique faces significant challenges when considering issues such as time and space complexity and non-unique parameters that are required for removing coincidental links. This paper presents an improved approach by the use of the popular fuzzy c-means method to reinforce the clustering of coincidental links in an emerging social network derived from spatiotemporal data. In particular, all links are organised into two groups: strong links or weak links, prior to the running of the filtering process. The efficacy of the modified version is demonstrated via systematic experimental comparisons against the performance of the original method.

Keywords Social networks · Coincidental links · Clustering algorithm · Fuzzy c-means · Spatiotemporal data

1 Introduction

Social network analysis has become a popular technique for many problem-solving applications. For instance, it has been increasingly applied to behavioural ecology, especially for animal behaviours analysis. Indeed, animal social networks have three distinct properties Krause et al. (2009):

1. Supporting the analysis of complex networks whose individuals have many features to consider.
2. Permitting the exploration of network structures at different levels such as individuals, dyad, group, and population, enabling the generation of larger networks from pairwise interactions.
3. Helping discover the influence of individual behaviours on a population and assess the fitness of population on individuals.

In study of behavioural ecology, consistent physical proximity of individuals is generally regard as a proxy for social links (Wilson 1975). For this, spatiotemporal data streams are required, which have been applied in animal behaviour analysis (e.g., Aebischer et al. 1993; White and Garrott 2012). Traditional approaches to constructing such animal social networks are based on a significant hypothesis named the Gambit of Group (GoG) Whitehead and Dufault (1999), which indicates that there are social connections between individuals which co-occur within a certain time period and at the same location. However, as reported in the literature (e.g., Psorakis et al. 2015; Franks et al. 2010; Croft et al. 2008; James et al. 2009; Whitehead 2008), this assumption often leads to important limitations, such as the resulting network containing many links that do not represent potential

Communicated by F. Chao, Q. Zhang.

✉ Qiang Shen
qqs@aber.ac.uk

Pu Zhang
puz@aber.ac.uk

[1] Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK

Published online: 09 July 2018

relationships or it over-counting associations in large groups. In particular, time windows are commonly employed to generate social networks by linking two individuals within the same time period Lauw et al. (2005). Yet, the size of an appropriate time window is difficult to determine, whereas it may have remarkable impact upon the outcomes of the final social network derived Krings et al. (2012).

Having taken notice of the problems of following the time window approach, the work in Psorakis et al. (2012) has presented an alternative to extract an animal social network from spatiotemporal data streams. The data streams applied in that research came from a part of an ongoing long-term field study of great tits regarding their foraging process, by setting different feeders in the forest during a number of winter seasons.

The procedure consists of three steps. First, the tits are clustered into different 'gathering events' according to their recorded time and location. Then, the tits which appear in the same gathering events are linked, forming a social network. In such a social network, there are many tits appearing in the same events by chance, which leads the network including coincidental links. Therefore, in the last step, a so-called null model is designed to filter those coincidental links. The null model works based on the presumption that all individuals covered by such a model have no direct relationship with each other and appear in gathering events randomly. Thus, all links generated in this case will be regarded as coincidental links. Any links with weaker strengths below these coincidental links can be filtered while retaining just the ones with greater strengths. Through experimental comparison with the time window method of Psorakis et al. (2015), this approach can better reconstruct the latent social network. This approach is also employed in Reynolds (2015), and the generated social networks can be analysed using solution mechanisms working for hidden Markov models.

While a useful approach, there are two significant problems in the use of null model: that both the time complexity and space complexity are generally very large and that the threshold used to perform filtering is not unique. To address these limitations, the popular fuzzy c-means algorithm as reported in Bezdek et al. (1984) is applied to modify the coincidental link filter. It clusters links into two groups: strong links and weak links, where weak links are regarded coincidental and strong links are used to build the final network. According to systematic experimental comparisons, the filter based on the use of fuzzy c-means has led to improved results over the use of the null model, while having lower time complexity.

The rest of this paper is structured as follows. Section 2 reviews the general generation process of animal social networks. Section 3 describes two coincidental link filtering methods, the conventional null model and the proposed one that utilises fuzzy c-means clustering. Section 4 details the setting of the experiments carried out and the results of comparative experimental evaluations. Finally, Sect. 5 concludes this paper with future research pointed out.

2 Network generation

This section introduces the general process of generating animal social networks between individuals from spatiotemporal data set. The underlying presumption for generating links is still based on the previously mentioned notion of GoG (Whitehead and Dufault 1999).

Presumption 1 *If individuals arrive in the same location at close time points, then they may be clustered into the same group. This process is called 'gathering event.' If two individuals appear in the same gathering event, then they can be linked together.*

An implication of this presumption is that locations are treated as independent of one another. This means that (sub-)networks may be built for each location of interest separately and then the emerging (sub-)networks can be integrated over all the locations to obtain the final network. In this regard, a social network is inferred exclusively by time records in the data stream.

To implement the above presumption, the following techniques have been developed, which will be detailed below:

1. Selection of arrival time records from the dataset.
2. Clustering of arrival time records into different gathering events.
3. Link-up of individuals if appearing in a common gathering event.

2.1 Arrival time record selection

In general, a spatiotemporal data stream may include a large number of records. However, only arrival time records are valuable for the network generation procedure, with the concept of arrival time as defined below:

Presumption 2 *If an individual is not recorded over a certain time period Δt, the next recorded time of the individual is regarded as the arrival time.*

To select an arrival time record from a large dataset, the time period Δt should be settled first. The size of Δt will influence the size of arrival time records and the accuracy of the subsequent gathering events. To obtain an appropriate Δt, a distribution of arrival time records is built up. A trial-and-error process is run to increase Δt until it clearly changes the distribution, and the point at which such a change takes place is returned as the maximum value of Δt. With the use of the

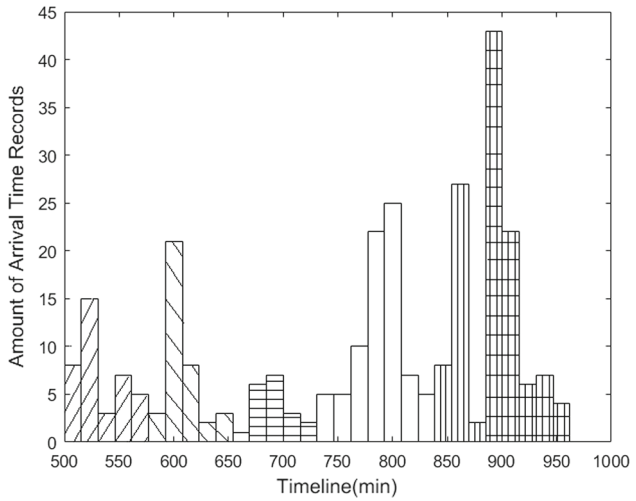

Fig. 1 An instance of arrival time records; as reflected by the distribution, records can be clustered into 6 groups that are termed 'gathering events'

maximum Δt, on the one hand, the accuracy of the process can be maintained, and on the other hand, the size of dataset is reduced which, in turn, also helps reduce the running time and storage space.

2.2 Gathering events clustering

Arrival time records are typically concentrated on several special periods as exemplified in Fig. 1. Such groups are termed 'gathering events.'

To cluster the records into different gathering events, Gaussian mixture model (GMM) Bilmes et al. (1998), a classical clustering algorithm is applied in the original work of Psorakis et al. (2012) and the popular EM algorithm Bilmes et al. (1998) is employed to generate it. The EM algorithm applied for the determination of GMM is outlined in Algorithm 1; a more detailed formal description of this algorithm is beyond the scope of this paper, but can be found in Dempster et al. (1977) and many of its follow-up work.

Algorithm 1: Applied EM algorithm

1. Initialisation: Calculate the initial means and variances of the Gaussian models using the global dataset statistics, with the number of the models fixed to a certain predefined value.
2. E-step (Expectation): For each data point, compute its conditional probability under each model using the current setup of the parameters.
3. M-step (Maximisation): For each Gaussian model, update its parameters using the data and their corresponding conditional probabilities.
4. Iteration: Iterate the above E- and M- steps until the likelihood converges or becomes smaller than a preset threshold.

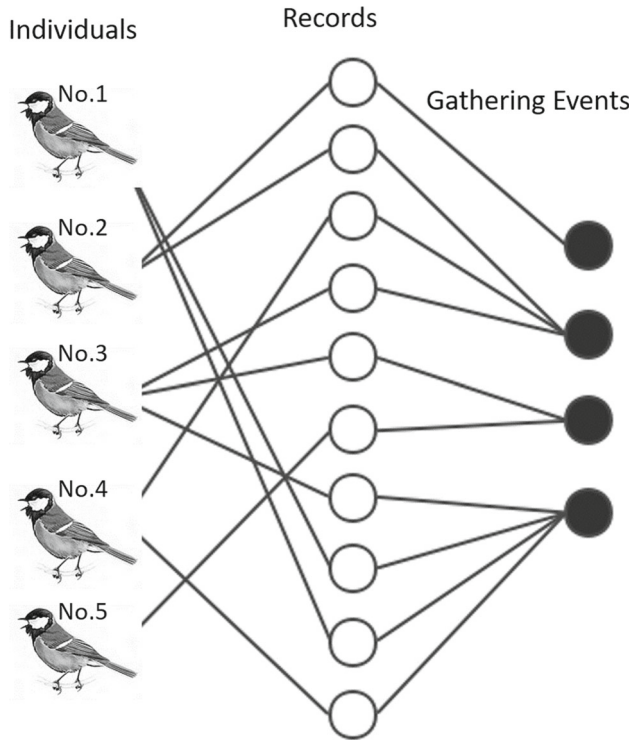

Fig. 2 Connection network consisting of two layers of nodes, with the left layer representing the records each individual involves and the right the gathering event each record belongs to

The result is described by a record-to-event matrix $RE \in R^{Z \times K}$, where Z is the number of arrival time records and K is the number of gathering events. Each record is only clustered into one gathering event while a single gathering event may include many records.

2.3 Link generation

According to Presumption 2, to identify links between individuals, the relationship between individuals and gathering events needs to be generated first. Since arrival time records have been clustered into gathering events (in the last step), while the identities of individuals have been included in the records, a connection network among individuals, records, and gathering events can be built up. Figure 2 shows a such network.

In a such network, most individuals may be involved in more than one record, and the records may belong to different gathering events. Thus, many individuals may attend at more than one gathering event. Based on that, the connections between individuals and events may be generated as shown in Fig 3. The result is described by an individual-to-event matrix $IE \in R^{N \times K}$, where N represents the number of individuals. Elements in this matrix represent the number of records an individual attends at a certain gathering event.

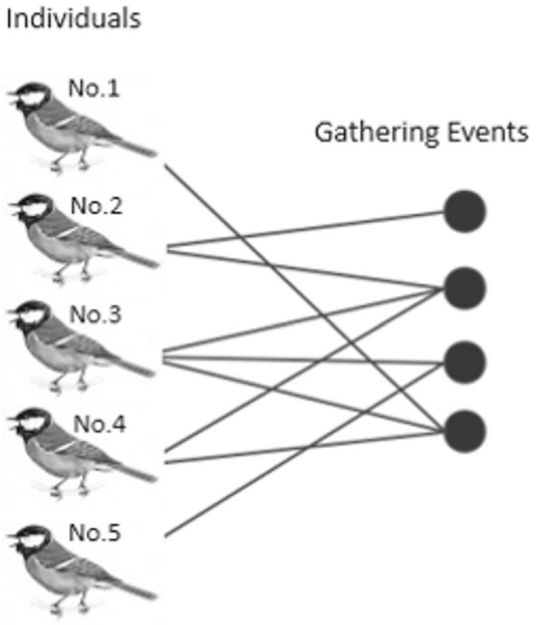

Fig. 3 Relationship between individuals and gathering events, representing which gathering events an individual appears in

Considering that different individuals may involve in a different number of records, the elements therefore need to be normalised with reference to the proportion of each event. These new elements are each called a 'preference,' which jointly lead to a new individual-to-preference (IP) matrix. Formally, the elements in this matrix IP are:

$$p_{ij} = \frac{r_{ij}}{\sum_{j=1}^{K} r_{ij}} \tag{1}$$

where i stands for an individual an individual, j for a gathering event, $r_{ij} \in IE$, and as indicated previously, K is the number of gathering events.

The strength of a link or relationship can be determined by the preference similarity. In particular, the link strength (or sometimes termed link weight in the literature) between individuals i and j is computed by the summation of the minimum preferences of these two individuals in the K gathering events:

$$a_{ij} = \sum_{k=1}^{K} \min(p_{ik}, p_{jk}) \tag{2}$$

Resulting strengths within the network can be concisely described by an adjacency matrix $NET \in R^{N \times N}$. Table 1 shows such a matrix generated from the data given in Fig. 3.

Table 1 Strength of links

	1	2	3	4	5
1	–	0	0.33	0.5	0
2	0	–	0.33	0.5	0
3	0.33	0.33	–	0.67	0.33
4	0.5	0.5	0.67	–	0.33
5	0	0	0.33	0.33	–

3 Coincidental link filtering

Coincidental links are a problem generally existing in social network modelling mainly due to spurious identification of (non-existing or very weak) links. It means that the linked individuals may appear together only by chance, or that the links with a low strength may exist in the matrix NET. Setting a threshold helps define normal links and filter out certain coincidental links. That is, links with a strength larger than a given threshold will be retained to become a link established in the learned network and those with a strength less than the threshold will be regarded as spurious ones and, therefore, filtered out. As an important additional benefit (apart than removing spurious links), coincidental filtering makes the structure of a generated network simpler, thereby making any subsequent reasoning computation simpler also.

3.1 Null model

A null model is one that is derived on the basis of the so-called null hypothesis (Moore et al. 2009). It assumes that all individuals so considered have no relationship with each other and the foraging process is totally random. All links generated in this case should then be regarded as coincidental links. Otherwise, any link whose strength as given in the matrix NET is larger than the threshold is retained as a final link. As such, the notion of null models is an important concept in animal social network construction based on the analysis of animal behaviours Farine and Whitehead (2015) Farine (2017).

Null model generation consists of two procedures: random process simulation and threshold selection. The random process simulation shuffles each row's elements in the IP matrix, generating a new matrix IP', which breaks up the original connections. Then, a new link matrix NET' can be produced (by the same process as described in Sect. 2.3). In order to minimise the generation of any coincidental result during the shuffling process, it is repeated T times to obtain T strengths for each link. However, in applications of this technique, the number of T needs to be decided on the basis of trial and error.

For the threshold selection procedure, statistical methods are applied to identify an appropriate threshold that represents a certain significance level α; all links with a strength less than this threshold will be regarded to be spurious and, hence, removed from the emerging network. There are two probability-based methods that may be applied for this: empirical distribution and normal distribution. Empirical distribution is cumulative, measuring the proportion of those data objects which are less than or equal to a specific value t. For the dataset $X = x_1, x_2, ..., x_n$, the distribution function $F(t)$ is defined by the following, as of Van der Vaart (2000):

$$F(t) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i \leq t) \tag{3}$$

From this, the threshold for empirical distribution is simply required to satisfy the following:

$$F(\text{threshold}) \geq \alpha. \tag{4}$$

Another method is based on normal distribution or Gaussian distribution, which is defined by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

where μ is the expectation and σ the standard deviation. For such a distribution, the threshold is set as $(\mu + n\sigma)$, which is associated with the given significance level α. Once the α is determined, the threshold is determined as well. From which the spurious links can be filtered out in a straightforward manner, by comparing the elements in the link strength matrix against this threshold value.

3.2 Problems with null model

Null model provides an approach to filtering coincidental links. However, there are two important limitations in using it:

1. **Computational complexity in time and space is large** According to the description of the null model, the time complexity is $O(NT^2)$, where N is the number of links and T is the number of shuffling times (which is typically a large number) that is required to obtain a set of strengths for each link. Empirically, T is larger than 50 to make the thresholds stable even for a moderately sized problem.Furthermore, in the null model, all T shuffled results (in terms of NET') have to be stored for subsequent threshold calculation. Thus, the space complexity is $O(NT)$.
2. **Thresholds may be different for different links** Differences between individuals may require different thresholds to be used, although this problem has been addressed previously (in Sect. 2.3) by normalising the IE matrix to the IP matrix. Nevertheless, it is difficult to extend the model given that multiple thresholds are often necessary to be employed because for any new link discovered, its threshold has to be calculated.

Therefore, in order to modify the existing approach for spurious link filtering while better exploiting the information contained within the links, clustering is herein applied to group the links into two distinct categories: strong links and weak links. The particular clustering algorithm employed in this work is fuzzy c-means (owing to its popularity, effectiveness and availability), which is outlined below.

3.3 Fuzzy C-means filter

The purpose of coincidental link filtering is to reduce spurious links. Through dividing links into two different groups, strong and weak, those weak links can be filtered out while retaining the strong ones. Unlike conventional clustering algorithms which only allow one cluster for each instance, fuzzy c-means allows for an instance to belong to different clusters with a different membership degree each.

Fuzzy c-means is originally developed in Dunn (1973) and subsequently improved in Bezdek et al. (1981). It clusters data by computing object distances to each emerging cluster centre. However, instead of using a Boolean distance metric directly as k-means (MacKay 2003), it calculates the membership of an object to decide on the clustering result. A minimum distance represents maximum membership. Here, the membership M is defined by

$$M = \left(\sum_{j=1}^{C}\left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|}\right)^n\right) - 1 \tag{6}$$

and the objective function (Bezdek et al. 1984) is defined by

$$J_m(U, V) = \sum_{k=1}^{N}\sum_{i=1}^{C}(u_{ik})^m \|x_k - v_i\|^2 \tag{7}$$

where x_k is a data object, m is the weighting exponent that controls the weight of each component, C is the number of cluster centres, N is the number of objects, V is the set of centres and $v_i \in V$, U is the set of membership functions and $u_{ik} \in U$ representing the membership of the object k belonging to the centre i, and $\|x_k - v_i\|$ represents the similarity between the object k and the centre i.

To minimise the objective function J_m, fuzzy c-means updates the membership function u_{ik} and the centre v_i itera-

tively by

$$v_i = \sum_{k=1}^{N} (u_{ik})^m x_k / \sum_{k=1}^{N} (u_{ik})^m \tag{8}$$

$$u_{ik} = \left(\sum_{j=1}^{C} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{2/m-1} \right)^{-1} \tag{9}$$

The procedure of the fuzzy c-means algorithm is summarised in Algorithm 2.

Algorithm 2: Fuzzy C-Means

1 Initialise $u_{ik} \in U^{(0)}$ randomly.
2 Set max iteration number L, termination condition ε.
3 Update centres $v_i \in V^{(k)}$ by $U^{(k)}$ and Equ. (8).
4 Update membership $u_{ik} \in U^{(k+1)}$ by $V^{(k)}$ and Equ. (9).
5 If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ or k+1 = L, then stop; otherwise, set $U^{(k)} = U^{(k+1)}$ and return to **step 3**.

Supported by the use of fuzzy c-means, the modified filtering process can be summarised as follows:

1. Link clustering: All links are categorised using fuzzy c-means into two groups according to their strength measures with respect to a predefined threshold (50% is used in this work), such that each strong link is associated with a larger (i.e., $\geq$ 50%) strength, while a weak link is associated with a smaller ($<$ 50%) strength measure.
2. Link filtering: Strong links are retained to form the final network while weak ones are filtered out.

Comparing with the null mode, in general, the time complexity of fuzzy c-means is $O(NCT)$, where N is the number of links, C is the number of link clusters and T is the number of iterations to run by the procedure. Herein, C is set to 2 since links are grouped into two categories. In comparison, the time complexity of null model is $O(NT^2)$ as mentioned previously. Note that the value of T for both methods will increase along with the increase in the size of the dataset. Thus, fuzzy c-means has lower time complexity. The real running times required by these two methods will be experimentally compared later.

The space complexity of running fuzzy c-means is $O(NC)$, where N is the number of links and C is the number of link clusters which is also set to 2 as indicated above. Within each iteration, fuzzy c-means is only required to update and store NC items. In comparison, running the null model, the space complexity is $O(NT)$ as mentioned previously. Note that the value of T is generally larger than 50, which is much larger than $C = 2$. Thus, fuzzy c-means also involves a lower space complexity.

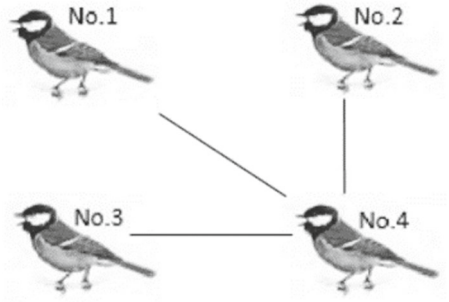

Fig. 4 Final social network filtered by fuzzy c-means filter

After applying the fuzzy c-means filter, the links in Table 1 will be divided into two categories. Links 1–4, 2–4, and 3–4 will be retained as strong links, while the other links will be filtered out as weak ones. In so doing, the final social network generated from the given example data is shown in Fig. 4.

4 Experimental evaluation

The experiments reported herein have two purposes. The first is to compare the performance of using two different clustering algorithms in implementing the link generation process, which are fuzzy c-means and Gaussian mixture model (Reynolds 2015). The second is to compare the performance of the modified method (with fuzzy c-means) and the original one (with null model) in coincidental link filtering.

To enable fair comparison, ground truth is ideally required to act as the golden standard for the networks to be built. However, there is no information regarding the ground truth available for the datasets used in the literature on animal social networks that are used to facilitate comparative studies. Blood relative is employed to evaluate the results in Psorakis et al. (2012), but only blood relation cannot represent all types of relationship and is not sufficient to serve as the ground truth.

In light of this observation, two different types of benchmark dataset are applied in the present investigation. One is based on the use of labelled datasets which are suitable for both classification and clustering problem. The underlying inference process remains the same. For coincidental link filtering, the ground truth for such datasets is utilised so that if two linked individuals have the same label, then this link is regarded as positive (i.e., the link is retained in the emerging network); otherwise, the link is negative (i.e., the link is removed). The other is generating artificial data streams with similar statistical properties as ground truth, given a fully-observed social network Psorakis et al. (2015). The inferred networks will be compared against the ground truth to evaluate their performance.

Again, for fair comparison, experiments are also carried out using the original dataset that was adopted in the original paper on animal social networks.

4.1 Labelled datasets

The construction process of the ground truth is similar to the inference process of the original work as reported in Psorakis et al. (2012), upon which this research is based. As indicated above, the structure of the employed datasets is also similar to that of the spatiotemporal datasets used there. The only difference is that unlike the dataset used in the original work where individuals can have similar memberships to multiple gathering events (as indicated in Sect. 2.3), the individuals in the datasets currently used can only have one high membership to a specific cluster. In other words, it can be regarded as a special condition of the original work when each individual mainly appears in one specific gathering event.

4.1.1 Datasets used

The following three data sets are adopted to perform this set of experiments: the iris dataset, the seeds dataset, and the wines dataset. For these labelled datasets, the number of positive links N can be calculated as follows:

$$N = \sum_{i=1}^{k} \frac{1}{2} n_i (n_i - 1) \tag{10}$$

where k is the number of classes and n_i is the number of objects belonging to class i. All these datasets contain features that are numerical. The following gives an overview of these datasets in terms of their properties.

Iris Dataset (Fisher 1936) consists of 150 instances, 4 features, and 3 classes. Each class includes 50 instances. There are 11175 links without involving any filtering process. According to Eq. (10), there are 3675 links that are positive, which jointly form the ground truth for testing.

Seed Dataset (Charytanowicz et al. 2010) consists of 210 instances, 7 features, and 3 classes. Each class includes 70 instances. There are 21945 links without involving any filtering process. According to Eq. (10), there are 7245 links that are positive, which jointly form the ground truth for testing.

Wines Dataset (Vandeginste 1990) consists of 178 instances, 13 features, and 3 classes. The three classes includes 71, 59, 48 instances, respectively. There are 15753 links without involving any filtering process. According to Eq. (10), there are 5324 links that are positive, which jointly form the ground truth for testing.

4.1.2 Experimental setup

Apart from the ground truth, there are a number of parameters that need to be set up in order to conduct the experiments. In particular, for the generation of gathering events, according to the given class labels, the number of gathering events (K) is set to 3. Since this parameter affects both fuzzy c-means and GMM, the performance of these two algorithms are experimentally compared also on the generation of gathering events.

For coincidental links filtering, in the original null model, the number of times of shuffling T is empirically set to 50, and 68% is employed as the significant level α (with normal distribution selected for threshold determination). In the implementation of the current work for the present experimental evaluation, the number of link clusters (C) is set to 2, which represents two link strength categories: strong and weak links.

The original and modified approaches are applied to generate and filter links and they are compared with regard to four aspects: the precision, the recall, the F1-score, and the running time of filtering process, where F1-score is a combination of precision and recall and is the major criterion to indicate the overall learning performance.

4.1.3 Results and discussion

The experimental results regarding the three datasets are displayed in Tables 2, 3, and 4, respectively. From these results, it can be seen that the coincidental filtering process can effectively reduce the total number of links while retaining the most valuable ones, thereby helping decrease storage space while increasing retrieve time significantly. Importantly, in terms of link generation, fuzzy c-means performs better than GMM while increasing the amount of domain features.

Looking into the results more specifically, it is revealed that for link filtering, fuzzy c-means based filters have a larger F1-score than that the null model based. This demonstrates from one aspect that the modified approach outperforms the original. In particular, FCM with $C = 2$ typically offers the best performance, retaining the most positive links with the highest recall rate. However, due to the largest amount of links it keeps, certain coincident links fail to be removed, which leads to a relatively low precision rate. On the contrary, the null model has a higher precision rate than the former. Yet, at the same time, certain positive links may be wrongly removed, leading to a relatively low recall rate and slightly worsening results than the filter based on FCM regarding the F1-score overall.

Table 2 Results of original and modified methods on iris dataset

Link generation	Link filter	Precision	Recall	F-1 score	Running time
FCM ($K = 3$)	FCM ($C = 2$)	0.7635	0.9284	**0.8379**	0.154
	Null model	0.82	0.7973	0.8085	3.444
GMM ($K = 3$)	FCM ($C = 2$)	0.9324	0.9390	**0.9357**	0.119
	Null model	0.9342	0.9268	0.9305	3.532

Table 3 Results of original and modified methods on seeds dataset

Link generation	Link filter	Precision	Recall	F-1 score	Running time
FCM ($K = 3$)	FCM ($C = 2$)	0.7423	0.8875	**0.8084**	0.407
	Null model	0.8286	0.7760	0.8014	12.335
GMM ($K = 3$)	FCM ($C = 2$)	0.7501	0.7619	**0.7560**	0.406
	Null model	0.7493	0.7469	0.7481	12.473

Table 4 Results of original and modified methods on wines dataset

Link generation	Link filter	Precision	Recall	F-1 score	Running time
FCM ($K = 3$)	FCM ($C = 2$)	0.7649	0.9673	**0.8543**	0.223
	Null model	0.9174	0.797	0.8530	6.476
GMM ($K = 3$)	FCM ($C = 2$)	0.7798	0.7878	**0.7838**	0.235
	Null model	0.7811	0.7795	0.7803	6.637

In terms of running time performance, the result tables clearly show that the modified method takes much less time than the original method. As such, these results empirically confirm that the time complexity of the modified approach is lower than that of the original (as indicated in Sect. 3.3). Furthermore, with the increase in the amount of instances the original method takes much longer time, while the modified approach as presented herein only raises the cost in computational time a little. Putting the F1-score and running time together, this set of experimental results collectively demonstrates that the employment of FCM (with $C = 2$) leads to an overall winner.

4.2 Artificial data streams

As mentioned above, ground-truth network structure is not available in real-world spatiotemporal data streams. To address this problem, a procedure as presented in Psorakis et al. (2015) is adopted to generate artificial data streams with similar statistical properties regarding a certain social network. A given network is considered as the ground truth network and is subsequently converted to data streams. The connected individuals are assumed to appear in temporal proximity. Thus, the social network inferred from the artificial data streams can have a ground truth to be compared against. This procedure is detailed in the supplementary material of Psorakis et al. (2015) and it is available online as a MATLAB script at: https://github.com/ipsorakis/GEgenerator.

In particular, two different artificial social networks are employed as ground truth networks in the experiments. One includes 50 individuals which are evenly divided into five groups. The converted temporal data stream includes 1225 records. The other includes 100 individuals which are evenly grouped into 10 categories, with 4950 records generated as data streams.

4.2.1 Experimental setup

Distinct from the above experiments with labelled datasets, the number of gathering events (K) in link generation is uncertain, which will affect the accuracy of the learned social networks. Therefore, the influence of K on both the original and modified method will also be experimentally compared in the following. In this set of experiments, Gaussian mixture model is applied in the link generation process since there is only one feature concerned herein, while fuzzy c-means with $C = 2$ is employed to compare against the use of the null model due to its efficacy, reflecting the observations over the previous results on labelled datasets. Both F1-score and running time are used as the evaluation criteria in the current experiments. Again, this reflects the previously achieved empirical results (see the preceding subsection),

4.2.2 Results and discussion

Table 5 presents the results of experimental comparison over the first artificial data stream. It can be seen from these results that FCM with $C = 2$ has a greater F1-score and less running

Table 5 Results of original and modified methods on artificial data streams with 50 individuals

Filter method	Evaluation	$K = 30$	$K = 40$	$K = 50$
FCM ($C = 2$)	F1-score	0.9336	0.9933	1.000
	Running time	0.0150	0.0300	0.0126
Null model	F1-score	0.9128	0.9202	0.9595
	Running time	0.2581	0.3437	0.2595

time than the null model, while both of them have a stable performance.

The results of application to the second data stream are summarised in Table 6. Contrary to the first data stream, there are significant differences between the performances of the two approaches compared. In particular, the results show that the number of clusters in link generation actually affects the filtering process and the efficacy of the final social networks.

To have a closer examination, consider a more detailed experimental process where the parameter K is set to vary. To start with, let K equal to 200 or a number around it, it is observed that the link generation processes of both original and modified approaches obtain the best results (as compared to the use of a lower K). Importantly, however, when K reduces, the modified approach still can generate stable results with a high accuracy and short running time while the performance of the original method falls down. This shows the robustness of the present work.

Note that in general, how to determine the number of clusters K in any clustering algorithm is a major topic in the area of data-driven clustering. Interestingly, the modified link filter implemented with fuzzy c-means can help extending the value range of K, making it less sensitive to the given problem while saving the computation required by the link generation process. Altogether, the modified approach significantly outperforms the original that utilises the null model.

4.3 Spatiotemporal dataset without ground truth

The dataset used in this experiment is the one adopted in the original work MacKay (2003), on animal social networks. It came from a large amount of research into the foraging process in a population of Parus major at Wytham Woods, near Oxford, the UK from 2007 to 2009 Psorakis et al. (2012). The dataset includes 1,032,797 records of 1241 different birds foraging in 69 different locations. Each record consists of three attributions: Bird ID, Time Stamp, and Location ID.

It is very difficult to display the full network because of the huge scale of the dataset. Therefore, as with the original work, 1000 records are randomly selected from the 1741 records to develop a network with both the original and the currently modified method for comparison. This process is independently carried out for 100 times. The outcome is that the

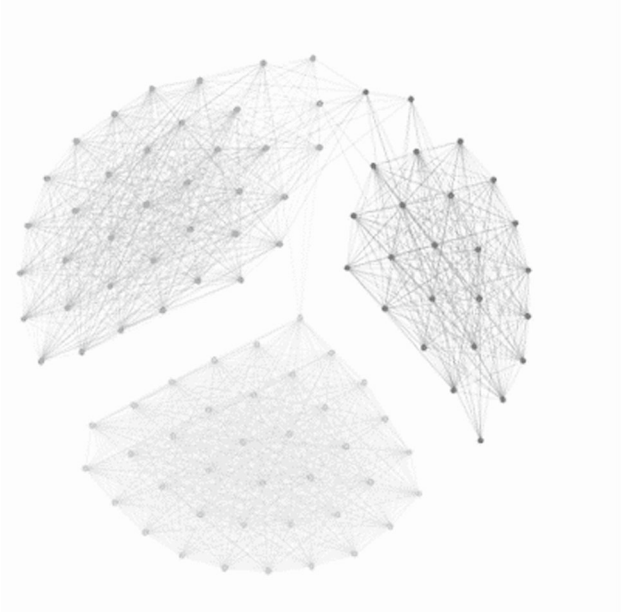

Fig. 5 Network generated from randomly selected 1000 records based on fuzzy c-means

modified method returns 1204 strong links on average with a standard deviation of 2.6842×10^4, while the original method retains 1249 links on average with a standard deviation of 8.0140×10^4. Interestingly, they have 1154 links in common (more than 90% of both methods). Evidently, the modified method is the one with stable performance. Moreover, the present work leads to 45 (or 3.6%) less links returned. Figure 5 showing the average result of running the modified method, and Fig. 6 doing that of the original method. Visually, given the complexity of the dataset, these two networks have resulted in similar structures.

For the entire dataset, the original method discovers 6857 strong links while the modified generates 6637 strong links. Between the two network, there are 6173 links in common. As such, the present approach allows a reduction of 684 (or just about 10%) links, significantly simplifying the resultant network structure. Once again, this result confirms that the modified method can indeed work to build more efficient animal social networks. However, for this particular dataset, there is no ground truth to perform a more detailed comparison of both methods in terms of what links have been removed. Nevertheless, previous results on datasets with ground truth have indicated that most (if not all) removed links from the network created by the original method (as of MacKay (2003)) are spurious ones.

5 Conclusion

This paper has introduced a method to infer animal social networks from spatiotemporal data streams with a modified

Table 6 Results of original and modified methods on artificial data streams with 100 individuals

Filter method	Evaluation	$K = 100$	$K = 150$	$K = 200$
FCM ($C = 2$)	F1-score	0.9912	0.9989	1.000
	Running time	0.0512	0.0882	0.0675
Null model	F1-score	0.7059	0.8257	0.9847
	Running time	1.5789	1.6748	1.7531

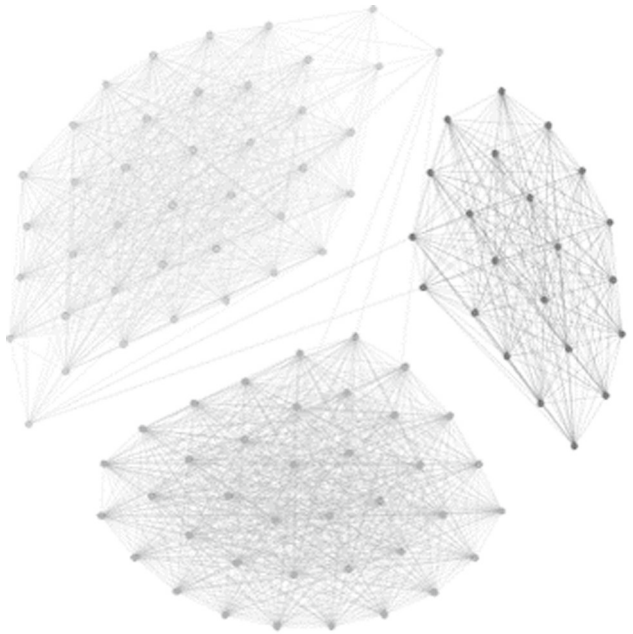

Fig. 6 Network generated from randomly selected 1000 records based on null model

procedure which enhances coincidental links filtering. The work has been motivated by the observation that the existing method clusters individuals into different gathering events and those belonging to the same gathering event are directly treated as linked, thereby producing many spurious links. To effectively filter coincidental links, fuzzy c-means has been applied to modifying this method by clustering the links into strong and weak links, thereby enabling the efficient removal of weak links. Systematic comparative experimented studies have demonstrated the effectiveness of this work. The analysis carried out in terms of the computational time and space complexity has shown the efficiency of this modified approach.

The presented procedure has its generality: It can be applied not only to animal social networks but also to human social networks, and the features addressed may be attributes different from just time and location. This has been shown to be feasible as datasets other than those relevant to animal social networks have also been utilised in the experiments carried out. The present work directly employs the original fuzzy c-means algorithm. However, a range of modified fuzzy c-means algorithms have been proposed in the literature, including kernel fuzzy c-means Zhang and Chen (2004) and suppressed fuzzy c-means Fan et al. (2003), for example. The approach proposed in this paper is sufficiently flexible, facilitating the use of such potential alternatives (though the employment of these more recent methods may incur more computational overheads).

The current work only deals with binary links within a social network. However, triple links may be inferred from such binary associations with the support of link analysis (Shen and Boongoen 2012; Su et al. 2013). This would be very useful in real-world settings where missing information regarding a third party needs to be inferred to enrich the social networks from neighbourhood binary relationships. This forms a major focus of further research. Also, instead of using fuzzy c-means or its immediate derivatives for coincidental link filtering, it is interesting to investigate whether performance may be reinforced if more advanced fuzzy clustering mechanisms (which are sufficiently efficient, e.g., Boongoen et al. 2011; Su et al. 2015 and Su et al. 2017) are used as an alternative.

Compliance with ethical standards

Funding: This study was mainly self-funded; other than the first author receiving a PhD scholarship from Aberystwyth University, no external funding was received in support of this research.

Conflict of interest: All authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

References

Aebischer NJ, Robertson PA, Kenward RE (1993) Compositional analysis of habitat use from animal radio-tracking data. Ecology 74(5):1313–1325

Bezdek JC, Coray C, Gunderson R, Watson J (1981) Detection and characterization of cluster substructure I. linear structure: fuzzy c-lines. SIAM J Appl Math 40(2):339–357

Bezdek JC, Ehrlich R, Full W (1984) Fcm: the fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203

Bilmes JA et al (1998) A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Int Comput Sci Inst 4(510):126

Boongoen T, Shang C, Iam-On N, Shen Q (2011) Extending data reliability measure to a filter approach for soft subspace clustering. IEEE Trans Syst Man Cybern Part B (Cybern) 41:1705–1714

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188

Charytanowicz M, Niewczas J, Kulczycki P, Kowalski PA, Łukasik S, Żak S (2010) Complete Gradient Clustering Algorithm for†Features Analysis of X-Ray Images. In: Pietka E, Kawa J (eds) Information Technologies in Biomedicine. Springer, Berlin, Heidelberg

Croft DP, James R, Krause J (2008) Exploring animal social networks. Princeton University Press, Princeton

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc 39:1–38

Dunn JC (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J Cybernetics 3(3):32–57

Fan J-L, Zhen W-Z, Xie W-X (2003) Suppressed fuzzy c-means clustering algorithm. Pattern Recognit Lett 24(9–10):1607–1612

Farine DR (2017) A guide to null models for animal social network analysis. Methods Ecol Evol 8(10):1309–1320

Farine DR, Whitehead H (2015) Constructing, conducting and interpreting animal social network analysis. J Anim Ecol 84(5):1144–1163

Franks DW, Ruxton GD, James R (2010) Sampling animal association networks with the gambit of the group. Behav Ecol Sociobiol 64(3):493–503

James R, Croft DP, Krause J (2009) Potential banana skins in animal social network analysis. Behav Ecol Sociobiol 63(7):989–997

Krause J, Lusseau D, James R (2009) Animal social networks: an introduction. Behav Ecol Sociobiol 63(7):967–973

Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. EPJ Data Sci 1(1):4

Lauw HW, Lim E-P, Pang H, Tan T-T (2005) Social network discovery by mining spatio-temporal events. Comput Math Org Theory 11(2):97–118

MacKay DJ (2003) Information theory, inference and learning algorithms. Cambridge University Press, Cambridge

Moore DS, McCabe GP, Craig BA (2009) Introduction to the practice of statistics. WH Freeman, New York

Psorakis I, Roberts SJ, Rezek I, Sheldon BC (2012) Inferring social network structure in ecological systems from spatio-temporal data streams. J R Soc Interface 9(76):3055–3066

Psorakis I, Voelkl B, Garroway CJ, Radersma R, Aplin LM, Crates RA, Culina A, Farine DR, Firth JA, Hinde CA et al (2015) Inferring social structure from temporal data. Behav Ecol Sociobiol 69(5):857–866

Reynolds D (2015) Gaussian mixture models. In: Li SZ, Jain AK (eds) Encyclopedia of biometrics. Springer, Boston, MA, pp 827–832

Shen Q, Boongoen T (2012) Fuzzy orders-of-magnitude-based link analysis for qualitative alias detection. IEEE Trans Knowl Data Eng 24(4):649–664

Su P, Shang C, Shen Q (2013) Link-based approach for bibliometric journal ranking. Soft Comput 17(12):2399–2410

Su P, Shang C, Shen Q (2015) A hierarchical fuzzy cluster ensemble approach and its application to big data clustering. J Intell Fuzzy Syst 28:2409–2421

Su P, Shang C, Shen Q (2017) Exploiting data reliability and fuzzy clustering for journal ranking. J Intell Fuzzy Syst 25(5):1306–1319

Van der Vaart AW (2000) Asymptotic statistics, vol 3. Cambridge university press

Vandeginste B (1990) PARVUS: An extendable package of programs for data exploration, classification and correlation, M. Forina, R. Leardi, C. Armanino and S. Lanteri, Elsevier, Amsterdam, 1988, Price: US $645 ISBN 0-444-43012-1. J Chemometrics 4(2):191–193

White GC, Garrott RA (2012) Analysis of wildlife radio-tracking data. Elsevier, London

Whitehead H (2008) Analyzing animal societies: quantitative methods for vertebrate social analysis. University of Chicago Press, Chicago

Whitehead H, Dufault S (1999) Techniques for analyzing vertebrate social structure using identified individuals. Adv Stud Behav 28:33–74

Wilson EO (1975) Sociobiology: the new synthesis. In: An anthology, philosophy of biology, p 339

Zhang D-Q, Chen S-C (2004) A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. Artif intell Med 32(1):37–50